

Combining Inclusion Polymorphism and Parametric Polymorphism

Sabine Glesner

Institut für Programmstrukturen
und Datenorganisation
Lehrstuhl Prof. Goos
Universität Karlsruhe
Postfach 6980, D 76128 Karlsruhe
Tel.: +49 / 721 - 608 7399
Fax: +49 / 721 - 300 47
glesner@ipd.info.uni-karlsruhe.de

Karl Stroetmann

Siemens AG
ZT SE 4

Otto-Hahn-Ring 6
D-81739 München
Tel.: +49 / 89 - 636 49 555
Fax: +49 / 89 - 636 42 284
Karl.Stroetmann@mchp.siemens.de

Abstract

We show that the question whether a term is typable is decidable for type systems combining inclusion polymorphism with parametric polymorphism provided the type constructors are at most unary. To prove this result we first reduce the typability problem to the problem of solving a system of type inequations. The result is then obtained by showing that the solvability of the resulting system of type inequations is decidable.

1 Introduction

As a common agreement, a flexible type system needs to contain inclusion as well as parametric polymorphism. Unfortunately, such a flexibility of the type system causes type inference to become hard or even undecidable. In this paper, we investigate the problem of checking the well-typedness of terms in the presence of inclusion polymorphism combined with parametric polymorphism. We show that in this case typability is decidable, provided that the type constructors are at most unary. This result has not been stated before.

The result of our paper can be used for the design of new type systems that combine both inclusion polymorphism and parametric polymorphism. Type systems of this kind are of interest for object-oriented programming languages. In particular, our result is applicable for the programming language Java which up to now does not allow for parametric polymorphism but will probably do so in future versions [MBL97, OW97].

Another area where our result is applicable is logic programming. A number of type systems have been designed in this area, e. g. [AM94, HT92, Pfe92, YFS92], but the systems that have been implemented so far either offer no inclusion polymorphism at all [HL94, SHC95] or impose stronger restrictions [Bei95b] than a type system that would be based on our result.

As it stands, our result cannot be applied to functional programming languages because these languages allow for the binary type constructor \rightarrow which takes two types σ and τ and returns the type $\sigma \rightarrow \tau$ of all functions mapping σ to τ .

In general, Tiuryn and Urzyczyn have shown that type inference for a type system which combines inclusion polymorphism and parametric polymorphism is undecidable for second-order types [TU96]. On the other hand, type inference for inclusion polymorphism combined with nullary type constructors is decidable: [Mit84, Mit91] presents an algorithm called MATCH, which solves type inequations in the case that only inequations between nullary type constructors are allowed. Fuh and Mishra [FM90] introduce a similar algorithm to solve the same problem. The logic programming language PROTOS-L [Bei95a] is based on this type system.

This paper is organized as follows: Section 2 contains a definition of the type language. In section 3 we define well-typed terms. Moreover, we show how the question whether a term is well-typed can be reduced to the problem of solving a system of type inequations. The solvability of these systems is shown to be decidable in section 4. Section 5 concludes.

2 Type Language

In this section we first introduce a language for describing types. Since types behave in many ways like terms, there is also a notion of *substitution*. This notion is defined in subsection 2.2.

2.1 Types

Types are constructed from type constructors and type parameters. The set of type constructors is partially ordered. This ordering is extended to types.

Definition 1 (Ordered Type Alphabet) An *ordered type alphabet* is a tuple $\mathfrak{A} = (\mathcal{A}, \#, \leq)$ such that

1. The *type alphabet* \mathcal{A} is a finite set of *type constructors*. Elements of \mathcal{A} are denoted by K, L, M, \dots .

2. (\mathcal{A}, \leq) is a partial order.
3. $\# : \mathcal{A} \rightarrow \mathbb{N}$ is a function assigning an *arity*, $\#K$, to every type constructor $K \in \mathcal{A}$. \diamond

Definition 2 (Types) To define *types*, we assume that an ordered type alphabet $\mathfrak{A} = (\mathcal{A}, \#, \leq)$ and a set $\mathfrak{P} = \{\alpha_i : i \in \mathbb{N}\}$ of *type parameters* are given. Then the set of *types* $\mathfrak{T} = \mathfrak{T}(\mathfrak{A}, \mathfrak{P})$ is defined inductively:

- $\alpha \in \mathfrak{T}$ for all $\alpha \in \mathfrak{P}$.
- If $K \in \mathcal{A}$, $\#K = n$, and $\sigma_i \in \mathfrak{T}$ for all $i = 1, \dots, n$, then $K(\sigma_1, \dots, \sigma_n) \in \mathfrak{T}$.

If $\#K = 0$, then we write K instead of $K()$. \diamond

Types are denoted by π , ϱ , σ , and τ , while parameters are denoted by α and β . A *monotype* is a type constructed without type parameters. If τ is a type, then $\text{Par}(\tau)$ denotes the set of type parameters in τ .

Next, we extend the relation \leq from \mathfrak{A} to the set of types $\mathfrak{T}(\mathfrak{A}, \mathfrak{P})$.

Definition 3 (Subtype Relation) Let $\mathfrak{A} = (\mathcal{A}, \#, \leq)$ be an ordered type alphabet and let $\mathfrak{T}(\mathfrak{A}, \mathfrak{P})$ be the set of types constructed from \mathfrak{A} . Then the subtype relation on $\mathfrak{T}(\mathfrak{A}, \mathfrak{P})$ is defined inductively:

1. If $\alpha \in \mathfrak{P}$, then $\alpha \leq \alpha$.
2. If $K, L \in \mathcal{A}$, $\#K = m$, and $\#L = n$, then $K(\sigma_1, \dots, \sigma_m) \leq L(\tau_1, \dots, \tau_n)$ holds iff $K \leq L$ and $\sigma_i \leq \tau_i$ for all $i = 1, \dots, \min(m, n)$. \diamond

Without further provisons, (\mathfrak{T}, \leq) is not a partial order. This is shown by the counter example given next.

Example 1 Assume that $\mathcal{A} = \{K_1, K_2, L_1, L_2\}$ where $\#K_i = 0$ and $\#L_i = 1$ for $i = 1, 2$. The ordering \leq on \mathcal{A} is defined by the following chain of inequations:

$$L_1 \leq K_1 \leq K_2 \leq L_2.$$

Then we have $L_1(K_2) \leq K_1$ and $K_1 \leq L_2(K_1)$. However, $L_1(K_2) \not\leq L_2(K_1)$. \diamond

This problem is caused by an incompatibility between the arity function $\# : \mathcal{A} \rightarrow \mathbb{N}$ and the ordering of the type alphabet.

Definition 4 (Compatible) Assume a type alphabet $\mathfrak{A} = (\mathcal{A}, \#, \leq)$ is given. Then the arity $\# : \mathcal{A} \rightarrow \mathbb{N}$ is *compatible* with the ordering \leq iff the following condition is satisfied for all type constructors K , L , and M :

$$K \leq L \wedge L \leq M \Rightarrow \min(\#K, \#M) \leq \#L. \quad \diamond$$

Convention: For the rest of this paper we assume the following: If an ordered type alphabet $(\mathcal{A}, \#, \leq)$ is given, then $\#$ is compatible with \leq .

Lemma 1 If $\mathfrak{A} = (\mathcal{A}, \#, \leq)$ is an ordered type alphabet, then $(\mathfrak{T}(\mathfrak{A}, \mathfrak{P}), \leq)$ is a partial order.

Proof: We need to show that the relation \leq is reflexive, antisymmetric, and transitive. In order to prove the reflexivity, we have to show $\sigma \leq \sigma$ for all types σ . This is done via a trivial induction on σ . To prove the antisymmetry, assume $\sigma \leq \tau$ and $\tau \leq \sigma$. We have to show $\sigma = \tau$. The proof proceeds by induction on σ .

1. σ is a parameter α . Because of $\sigma \leq \tau$ we know that $\tau = \alpha$.
2. $\sigma = L(\sigma_1, \dots, \sigma_l)$. Then $\tau = M(\tau_1, \dots, \tau_m)$ and we must have $L \leq M$ and $M \leq L$. Since \leq is a partial order on \mathcal{A} , we have $L = M$ and $l = m$. Further, we have

$$\begin{aligned} \sigma_i &\leq \tau_i \quad \text{for all } i = 1, \dots, l, \quad \text{and} \\ \tau_i &\leq \sigma_i \quad \text{for all } i = 1, \dots, l. \end{aligned}$$

The induction hypothesis yields $\sigma_i = \tau_i$ for all $i = 1, \dots, l$ and then $\sigma = \tau$ is immediate.

To prove the transitivity, assume that $\varrho, \sigma, \tau \in \mathfrak{T}(\mathfrak{A}, \mathfrak{P})$ are given such that $\varrho \leq \sigma$ and $\sigma \leq \tau$. We need to prove $\varrho \leq \tau$. The proof proceeds by induction on σ .

1. σ is a parameter α . Then ϱ is α and, similarly, τ is α . Obviously, $\varrho \leq \tau$.
2. σ is $L(\sigma_1, \dots, \sigma_l)$. Then $\varrho = K(\varrho_1, \dots, \varrho_k)$ and $\tau = M(\tau_1, \dots, \tau_m)$. The assumption $\varrho \leq \sigma$ yields $K \leq L$ and

$$\varrho_i \leq \sigma_i \quad \text{for all } i = 1, \dots, \min(k, l)$$

and, similarly, the assumption $\sigma \leq \tau$ yields $L \leq M$ and

$$\sigma_i \leq \tau_i \quad \text{for all } i = 1, \dots, \min(l, m).$$

Since \leq is a partial order on \mathcal{A} , we have $K \leq M$. Further, the induction hypothesis shows that

$$\varrho_i \leq \tau_i \quad \text{for all } i = 1, \dots, \min(k, l, m).$$

Since the arity $\#$ is compatible with \leq , we have $\min(k, m) \leq l$. Therefore, $\min(k, l, m) = \min(k, m)$. But then $\varrho \leq \tau$ is immediate. \square

2.2 Parameter Substitutions

Types behave in many ways like terms. Therefore there is also a notion of *substitution*. Since type parameters are substituted rather than variables, these substitutions are called *parameter substitutions*. Parameter substitutions are denoted by the capital Greek letters Θ , Φ , and Ψ .

Definition 5 (Parameter Substitution) A *parameter substitution* Θ is a finite set of pairs of the form

$$[\alpha_1 \mapsto \tau_1, \dots, \alpha_n \mapsto \tau_n]$$

where $\alpha_1, \dots, \alpha_n$ are distinct parameters and τ_1, \dots, τ_n are types. It is interpreted as a function mapping type parameters to types:

$$\Theta(\alpha) := \begin{cases} \tau_i & \text{if } \alpha = \alpha_i; \\ \alpha & \text{otherwise.} \end{cases}$$

This function is extended to types homomorphically:

$$\Theta(F(\sigma_1, \dots, \sigma_n)) := F(\Theta(\sigma_1), \dots, \Theta(\sigma_n)).$$

We use a postfix notation to denote the result of evaluating Θ on a type τ , i.e. we write $\tau\Theta$ instead of $\Theta(\tau)$.

The *domain* of Θ is defined as $\text{dom}(\Theta) := \{\alpha \mid \alpha \neq \alpha\Theta\}$. The set of parameters appearing in the range of a parameter substitution Φ is defined as

$$\text{Par}(\Phi) := \bigcup \{\text{Par}(\alpha\Phi) \mid \alpha \in \text{dom}(\Phi)\}.$$

A parameter substitution is called a *parameter renaming* iff it has the form

$$[\alpha_1 \mapsto \alpha_{\pi(1)}, \dots, \alpha_n \mapsto \alpha_{\pi(n)}]$$

where π is a permutation of the set $\{1, \dots, n\}$.

If Θ_1 and Θ_2 are parameter substitutions, then their *composition* $\Theta_1 \circ \Theta_2$ is defined such that $\alpha(\Theta_1 \circ \Theta_2) = (\alpha\Theta_1)\Theta_2$ holds for all type parameters α . \diamond

Parameter substitutions respect the ordering \leq on \mathfrak{T} .

Lemma 2 If Θ is a parameter substitution and $\sigma, \tau \in \mathfrak{T}$, then

$$\sigma \leq \tau \Rightarrow \sigma\Theta \leq \tau\Theta.$$

Proof: The proof is done by an induction following the definition of $\sigma \leq \tau$.

1. The case $\alpha \leq \alpha$ is obvious.
2. If $\sigma = K(\sigma_1, \dots, \sigma_m) \leq L(\tau_1, \dots, \tau_n) = \tau$, then $K \leq L$ and $\sigma_i \leq \tau_i$ for $i = 1, \dots, \min(m, n)$. Using the induction hypothesis we have $\sigma_i\Theta \leq \tau_i\Theta$ for all relevant i . Therefore $K(\sigma_1\Theta, \dots, \sigma_m\Theta) \leq L(\tau_1\Theta, \dots, \tau_n\Theta)$. \diamond

3 Well-Typed Terms

We define the set of well-typed terms in the first subsection. Then in subsection 3.2 we reduce the question whether a term is well-typed to the solvability of a system of type inequations.

3.1 Definition of Well-Typed Terms

We assume a set of functions symbols \mathcal{F} and a set of variables \mathcal{V} to be given. Every function symbol $f \in \mathcal{F}$ is supposed to have an *arity*.

Definition 6 (Terms) The set of terms $\mathcal{T}(\mathcal{F}, \mathcal{V})$ is defined inductively:

1. If $v \in \mathcal{V}$, then $v \in \mathcal{T}(\mathcal{F}, \mathcal{V})$.
2. If $f \in \Sigma$, f is n -ary, and $t_1, \dots, t_n \in \mathcal{T}(\mathcal{F}, \mathcal{V})$, then $f(t_1, \dots, t_n) \in \mathcal{T}(\mathcal{F}, \mathcal{V})$.

The set of variables occurring in a term t is defined by an obvious inductive definition and denoted by $\text{Var}(t)$. If this set is empty, then t is called a *closed term*. The set of closed terms is denoted by $\mathcal{T}(\mathcal{F})$. \diamond

Definition 7 (Signature) If f is n -ary, then its *signature* is a string of $n+1$ types. If $\sigma_1 \dots \sigma_n \tau$ is the signature of f , then this is communicated by writing

$$f : \sigma_1 \times \dots \times \sigma_n \rightarrow \tau.$$

In the following, we assume that every function symbol f has a signature.

A signature $\varrho_1 \dots \varrho_n \pi$ is *appropriate* for a function symbol f iff

$$f : \sigma_1 \times \dots \times \sigma_n \rightarrow \tau$$

and there exists a parameter substitution Θ such that $\pi = \tau\Theta$ and $\varrho_i = \sigma_i\Theta$ for $i = 1, \dots, n$. \diamond

Definition 8 (Type Assignment) A *type annotation* is a pair written as $t : \tau$ where t is a term and τ is a type. The type annotation $t : \tau$ is called a *variable annotation* if t is a variable. If $\Gamma = \{x_1 : \tau_1, \dots, x_n : \tau_n\}$ is a finite set of variable annotations such that the variables x_i are pairwise distinct, then we call Γ a *type*

assignment. If $\Gamma = \{x_1 : \tau_1, \dots, x_n : \tau_n\}$ is a type assignment, then we regard Γ as a function with domain $\{x_1, \dots, x_n\}$ mapping the variables x_i to the types τ_i , i.e. we have $\Gamma(x_i) = \tau_i$ for $i = 1, \dots, n$ and $\text{dom}(\Gamma) = \{x_1, \dots, x_n\}$. \diamond

Definition 9 (Well-Typed Term) The notion of a *well-typed* term is defined via a binary relation \vdash taking as its first argument a type assignment and as its second argument a type annotation. The definition of \vdash is done inductively:

1. If $\Gamma(x) \leq \pi$, then

$$\Gamma \vdash x : \pi.$$

2. If we have

- (a) $\Gamma \vdash s_i : \varrho_i$ for all $i = 1, \dots, n$,
- (b) $\sigma_1 \times \dots \times \sigma_n \rightarrow \tau$ is appropriate for f ,
- (c) $\varrho_i \leq \sigma_i$ for all $i = 1, \dots, n$, and
- (d) $\tau \leq \pi$,

then $\Gamma \vdash f(s_1, \dots, s_n) : \pi$.

A term t is *well-typed* iff there exist a type assignment Γ and a type τ such that $\Gamma \vdash t : \tau$. We read $\Gamma \vdash t : \tau$ as “ Γ entails $t : \tau$ ”. We call $\Gamma \vdash t : \tau$ a *type judgement*. \diamond

3.2 Type Checking

In this subsection, we reduce the question whether a term is well-typed to the solvability of a system of type inequations. Here, a *type inequation* is a pair of types written as $\sigma \preceq \tau$. A parameter substitution Θ *solves* a type inequation $\sigma \preceq \tau$ (denoted $\Theta \models \sigma \preceq \tau$) if $\sigma\Theta \leq \tau\Theta$. A *system* of type inequations is a set of type inequations. A parameter substitution Θ *solves* a system of type inequations \mathcal{I} (denoted $\Theta \models \mathcal{I}$) iff Θ solves every type inequation in \mathcal{I} .

Assume that Γ is a type assignment and $t : \tau$ is a type annotation such that $\text{Var}(t) \subseteq \text{dom}(\Gamma)$. We define a function $\text{ineq}(\Gamma, t : \tau)$ by induction on t such that $\text{ineq}(\Gamma, t : \tau)$ is a system of type inequations. A parameter substitution Θ will solve $\text{ineq}(\Gamma, t : \tau)$ iff $\Gamma\Theta \vdash t : \tau\Theta$. The inductive definition of $\text{ineq}(\Gamma, t : \tau)$ is given as follows:

1. $\text{ineq}(\Gamma, x : \tau) := \{\Gamma(x) \preceq \tau\}$
2. Assume the signature of f is given as $f : \sigma_1 \times \dots \times \sigma_n \rightarrow \sigma$, where the type parameters have been appropriately renamed so that they are *new*, i.e. the new parameters may occur neither in Γ nor in τ nor in any of the signatures used to construct $\text{ineq}(\Gamma, s_i : \sigma_i)$ for some $i = 1, \dots, n$. Then

$$\text{ineq}(\Gamma, f(s_1, \dots, s_n) : \tau) := \{\sigma \preceq \tau\} \cup \bigcup_{i=1}^n \text{ineq}(\Gamma, s_i : \sigma_i).$$

Before starting with the proofs of the soundness and completeness for the above transformation, we state some definitions: If Γ is a type assignment and $t : \tau$ is a type annotation, then $\Gamma \triangleright t : \tau$ is called a *hypothetical type judgement*. A parameter substitution Θ *solves* a hypothetical type judgement $\Gamma \triangleright t : \tau$ iff $\Gamma\Theta \vdash t : \tau\Theta$ holds. A *type constraint* is either a type inequation or a hypothetical type judgement. A parameter substitution Θ *solves* a set of type constraints C iff it solves every type inequation and every hypothetical type judgement in C . This is written $\Theta \models C$. We define a rewrite relation on sets of type constraints. It is the least transitive relation \rightsquigarrow such that:

1. $C \cup \{\Gamma \triangleright x : \tau\} \rightsquigarrow C \cup \{\Gamma(x) \preceq \tau\}$
2. Assume that the signature of f is given as $f : \sigma_1 \times \dots \times \sigma_n \rightarrow \sigma$ where the type parameters have been appropriately renamed so that they are new. Then

$$C \cup \{\Gamma \triangleright f(s_1, \dots, s_n) : \tau\} \rightsquigarrow C \cup \{\sigma \preceq \tau\} \cup \bigcup_{i=1}^n \{\Gamma \triangleright s_i : \sigma_i\}.$$

If a hypothetical type judgement $\Gamma \triangleright t : \tau$ is given, then the two rewrite rules can be used repeatedly until the set $\text{ineq}(\Gamma, t : \tau)$ is derived. This is easily seen by induction on t . Furthermore, the rewrite relation \rightsquigarrow satisfies the following invariants:

$$1. (\Theta \models C_2) \wedge (C_1 \rightsquigarrow C_2) \Rightarrow (\Theta \models C_1) \quad (\text{I}_1)$$

$$2. (\Theta \models C_1) \wedge (C_1 \rightsquigarrow C_2) \Rightarrow \exists \Psi. (\Theta \subseteq \Psi \wedge \Psi \models C_2) \quad (\text{I}_2)$$

Before proving these invariants, we show that they suffice to verify the soundness and completeness of our transformation.

Theorem 1 (Soundness of the Transformation) Assume Γ is a type assignment and $t : \tau$ is a type annotation. If $\Theta \models \text{ineq}(\Gamma, t : \tau)$, then $\Gamma\Theta \vdash t : \tau\Theta$.

Proof: Since the assumption is $\Theta \models \text{ineq}(\Gamma, t : \tau)$ and we know that $\{\Gamma \triangleright t : \tau\} \rightsquigarrow \text{ineq}(\Gamma, t : \tau)$, the invariant (I₁) shows that $\Theta \models \{\Gamma \triangleright t : \tau\}$. By definition, this implies $\Gamma\Theta \vdash t : \tau\Theta$. \square

Theorem 2 (Completeness of the Transformation) Assume Γ is a type assignment, $t : \tau$ is a type annotation, and Θ is a parameter substitution such that $\Gamma\Theta \vdash t : \tau\Theta$. Then, Θ can be extended to a parameter substitution Φ that is a solution of $\text{ineq}(\Gamma, t : \tau)$.

Proof: $\Gamma\Theta \vdash t : \tau\Theta$ implies $\Theta \models \Gamma \triangleright t : \tau$. Since $\{\Gamma \triangleright t : \tau\} \rightsquigarrow \text{ineq}(\Gamma, t : \tau)$, the invariant (I₂) shows that Θ can be extended to a parameter substitution Φ such that $\Phi \models \text{ineq}(\Gamma, t : \tau)$. \square

Proof of (I₁): According to the definition of the rewrite relation \rightsquigarrow , it suffices to consider the following two cases:

1. $C_1 = C \cup \{\Gamma \triangleright x : \tau\} \rightsquigarrow C \cup \{\Gamma(x) \preceq \tau\} = C_2$. The assumption is that $\Theta \models C_2$. Then $\Theta \models C$ and $\Gamma(x)\Theta \preceq \tau\Theta$. Therefore, $\Gamma\Theta \vdash x : \tau\Theta$ showing $\Theta \models C_1$.
2. $C_1 = C \cup \{\Gamma \triangleright f(s_1, \dots, s_n) : \tau\} \rightsquigarrow C \cup \{\sigma \preceq \tau\} \cup \bigcup_{i=1}^n \{\Gamma \triangleright s_i : \sigma_i\} = C_2$, where $f : \sigma_1 \times \dots \times \sigma_n \rightarrow \sigma$. According to the assumption, we have $\Theta \models C$, $\sigma\Theta \preceq \tau\Theta$, and $\Theta \models \Gamma \triangleright s_i : \sigma_i$ for $i = 1, \dots, n$. Then $\Gamma\Theta \vdash s_i : \sigma_i\Theta$ for $i = 1, \dots, n$. Therefore, $\Gamma\Theta \vdash f(s_1, \dots, s_n) : \tau\Theta$ and that yields the claim. \square

To prove the invariant (I₂) we need the following lemma, which follows directly from Defs. 7 and 9.

Lemma 3 Suppose that $t = f(s_1, \dots, s_n)$ and $f : \sigma_1 \times \dots \times \sigma_n \rightarrow \sigma$. Then $\Gamma \vdash t : \tau$ iff there is a parameter substitution Θ such that $\sigma\Theta \preceq \tau$ and $\Gamma \vdash s_i : \sigma_i\Theta$ for all $i = 1, \dots, n$.

Proof of (I₂): Again, it suffices to consider the following two cases corresponding to the definition of the relation \rightsquigarrow :

1. $C_1 = C \cup \{\Gamma \triangleright x : \tau\} \rightsquigarrow C \cup \{\Gamma(x) \preceq \tau\} = C_2$. The assumption is that $\Theta \models C_1$. Then $\Theta \models C$ and $\Gamma\Theta \vdash x : \tau\Theta$. Therefore, $\Gamma(x)\Theta \preceq \tau\Theta$. Define $\Psi := \Theta$.

2. $C_1 = C \cup \{\Gamma \triangleright f(s_1, \dots, s_n) : \tau\} \rightsquigarrow C \cup \{\sigma \preceq \tau\} \cup \bigcup_{i=1}^n \{\Gamma \triangleright s_i : \sigma_i\} = C_2$,
 where $f : \sigma_1 \times \dots \times \sigma_n \rightarrow \sigma$. W.l.o.g. we assume that the type parameters occurring in this signature do not occur in $\text{dom}(\Theta)$, since the type parameters in the signature can be renamed. According to our assumption, we have $\Theta \models C$ and $\Theta \models \{\Gamma \triangleright f(s_1, \dots, s_n) : \tau\}$. The latter implies $\Gamma\Theta \vdash f(s_1, \dots, s_n) : \tau\Theta$. Lemma 3 shows that there is a parameter substitution Φ such that $\Gamma\Theta \vdash s_i : \sigma_i\Phi$ for all $i = 1, \dots, n$ and $\sigma\Phi \leq \tau\Theta$. We can assume that $\text{dom}(\Phi)$ contains only type parameters occurring in the signature of f . Then $\text{dom}(\Theta) \cap \text{dom}(\Phi) = \emptyset$. Define $\Psi := \Theta \cup \Phi$. \square

When checking whether a term t is well-typed we want to compute a type assignment Γ and a type τ such that $\Gamma \vdash t : \tau$ holds. To this end, we define a most general type assignment Γ_{init} and a most general type τ_{init} : Let $\text{Var}(t)$ be the variables in t . Define $\Gamma_{\text{init}} = \bigcup_{x \in \text{Var}(t)} \{x : \alpha_x\}$ and $\tau_{\text{init}} = \alpha$ where α_x and α are distinct new type parameters. The claim now is that t is well-typed if and only if the set of type constraints $\text{ineq}(\Gamma_{\text{init}}, t : \tau_{\text{init}})$ is solvable.

Proof: “ \Rightarrow ”: Assume t is well-typed. Then there exists a type assignment Γ and a type τ such that $\Gamma \vdash t : \tau$. Define a parameter substitution Θ by setting $\Theta(\alpha_x) = \Gamma(x)$ for $x \in \text{Var}(t)$ and $\Theta(\alpha) = \tau$. Then we have $\Gamma(x) = \Gamma_{\text{init}}(x)\Theta$ and $\tau = \tau_{\text{init}}\Theta$ and therefore $\Theta \models \{\Gamma_{\text{init}} \triangleright t : \tau_{\text{init}}\}$. Since

$$\{\Gamma_{\text{init}} \triangleright t : \tau\} \rightsquigarrow \text{ineq}(\Gamma_{\text{init}}, t : \tau_{\text{init}}),$$

the invariant (I_2) shows that there exists a parameter substitution Ψ such that

$$\Psi \models \text{ineq}(\Gamma_{\text{init}}, t : \tau_{\text{init}}).$$

“ \Leftarrow ”: On the other hand, if $\Psi \models \text{ineq}(\Gamma_{\text{init}}, t : \tau_{\text{init}})$, then Theorem 1 shows that $\Gamma_{\text{init}}\Psi \vdash t : \tau_{\text{init}}\Psi$ holds. \square

Therefore, the problem whether a term t is well-typed is reduced to the problem of solving systems of type inequations.

4 Solving Systems of Type Inequations

In this section, we assume that type constructors are at most unary, i.e., given an ordered type alphabet $\mathfrak{A} = (\mathcal{A}, \#, \leq)$ we have that $\#K \leq 1$ for all $K \in \mathcal{A}$. We show that then it is decidable whether a system \mathcal{S} of type inequations is solvable. To this end we present an algorithm which effectively tests all possible instantiations for the type parameters in the type inequations. The fact that the type constructors are at most unary enables us to guarantee three important properties during this instantiation process: We do not create any additional parameters; we do not increase the overall number of inequations; and the depth of the terms in the type inequations does not increase. Therefore we can generate only finitely many systems of instantiated type inequations. If one of these systems is solvable, then we can construct a solution for \mathcal{S} .

4.1 Some Definitions

We start with some definitions necessary to formulate the algorithm for checking the solvability of systems of type inequations.

4.1.1 Solvability and Equivalence of Type Inequations

A system of type inequations \mathcal{I} is solvable (denoted $\Diamond\mathcal{I}$) iff there is a parameter substitution Φ such that $\Phi \models \mathcal{I}$. Two type inequations I_1 and I_2 are *equivalent* (denoted $I_1 \approx I_2$) iff a parameter substitution Φ solves I_1 if and only if Φ solves I_2 :

$$I_1 \approx I_2 \stackrel{\text{def}}{\iff} \forall \Phi \cdot (\Phi \models I_1 \iff \Phi \models I_2)$$

A type inequation I is equivalent to **true** (denoted $I \approx \mathbf{true}$) iff every parameter substitution solves I , it is equivalent to **false** (denoted $I \approx \mathbf{false}$) iff no parameter substitution solves I . Two systems of type inequations \mathcal{I}_1 and \mathcal{I}_2 are *equivalent* (denoted $\mathcal{I}_1 \approx \mathcal{I}_2$) iff a parameter substitution Φ solves \mathcal{I}_1 if and only if Φ solves \mathcal{I}_2 :

$$\mathcal{I}_1 \approx \mathcal{I}_2 \stackrel{\text{def}}{\iff} \forall \Phi \cdot (\Phi \models \mathcal{I}_1 \iff \Phi \models \mathcal{I}_2)$$

Next, a system of type inequations \mathcal{I} is equivalent to a set of systems of type inequations \mathcal{J} (denoted $\mathcal{I} \approx \mathcal{J}$) iff \mathcal{I} is solvable if and only if there is a system $\mathcal{J} \in \mathcal{J}$ such that \mathcal{J} is solvable:

$$\mathcal{I} \approx \mathcal{J} \stackrel{\text{def}}{\iff} (\diamond \mathcal{I} \iff \exists \mathcal{J} \in \mathcal{J} \cdot \diamond \mathcal{J}).$$

To proceed, we define the depth of a type inductively:

1. $\text{depth}(\alpha) := 0$ for all type parameters α .
2. $\text{depth}(K) := 1$ for all nullary type constructors K .
3. $\text{depth}(K(\sigma)) := 1 + \text{depth}(\sigma)$.

The depth of a type inequation is defined by taking the maximum:

$$\text{depth}(\sigma \preceq \tau) := \max(\text{depth}(\sigma), \text{depth}(\tau)).$$

Furthermore, we define $\text{depth}(\mathbf{true}) := \text{depth}(\mathbf{false}) := 0$. The function depth is then extended to systems of type inequations:

$$\text{depth}(\mathcal{I}) := \max\{\text{depth}(I) \mid I \in \mathcal{I}\}.$$

The depth of a parameter substitution Φ is defined as

$$\text{depth}(\Phi) := \max\{\text{depth}(\alpha\Phi) \mid \alpha \in \text{dom}(\Phi)\}.$$

We define the depth of the empty parameter substitution as 0. A system of inequations \mathcal{I} is *solvable at depth k* (denoted $\diamond_k \mathcal{I}$) iff there is a closed parameter substitution Φ such that $\Phi \models \mathcal{I}$ and $\text{depth}(\Phi) \leq k$.

4.1.2 Definition of nf

The function nf takes a type inequation as input and either produces an equivalent type inequation or yields **true** or **false**. The function is defined inductively.

1. $nf(\alpha \preceq \sigma) := \alpha \preceq \sigma$ and $nf(\sigma \preceq \alpha) := \sigma \preceq \alpha$ for every type parameter α .
2. $nf(K \preceq L) := \begin{cases} \mathbf{true} & \text{iff } K \leq L; \\ \mathbf{false} & \text{else.} \end{cases}$
3. $nf(K \preceq L(\tau)) := \begin{cases} \mathbf{true} & \text{iff } K \leq L; \\ \mathbf{false} & \text{else.} \end{cases}$
4. $nf(K(\sigma) \preceq L) := \begin{cases} \mathbf{true} & \text{iff } K \leq L; \\ \mathbf{false} & \text{else.} \end{cases}$
5. $nf(K(\sigma) \preceq L(\tau)) := \begin{cases} nf(\sigma \preceq \tau) & \text{iff } K \leq L; \\ \mathbf{false} & \text{else.} \end{cases} \quad \diamond$

It is easy to see that $nf(I) \approx I$ holds for every inequation I . We extend the function nf to systems of type inequations. First, we define an auxiliary function nf_{aux} :

$$nf_{\text{aux}}(\mathcal{I}) := \{nf(I) \mid I \in \mathcal{I} \wedge nf(I) \neq \text{true}\}.$$

Then, the function $nf(\mathcal{I})$ is defined as

$$nf(\mathcal{I}) := \begin{cases} \{\text{false}\} & \text{if } \text{false} \in nf_{\text{aux}}(\mathcal{I}); \\ nf_{\text{aux}}(\mathcal{I}) & \text{otherwise.} \end{cases}$$

It is easy to see that $\mathcal{I} \approx nf(\mathcal{I})$ for any system of type inequations \mathcal{I} .

4.1.3 Definition of *AllParSubst*

Next, we define the function *AllParSubst*. The input to *AllParSubst* is a finite set A of type parameters. The output is the set of parameter substitutions Φ such that $\text{dom}(\Phi) \subseteq A$, $\text{depth}(\Phi) \leq 1$, and $\text{Par}(\alpha\Phi) \subseteq \{\alpha\}$ but $\alpha\Phi \neq \alpha$ for all $\alpha \in A$. Therefore, $\text{AllParSubst}(A)$ is equal to the set

$$\{\Phi \mid \text{dom}(\Phi) \subseteq A \wedge \text{depth}(\Phi) \leq 1 \wedge \forall \alpha \in A. \text{Par}(\alpha\Phi) \subseteq \{\alpha\} \wedge \alpha\Phi \neq \alpha\}.$$

The function *AllParSubst* has the following properties:

1. *AllParSubst*(A) is finite.
This is true because the type alphabet is assumed to be finite. Therefore, given a finite set A of type parameter there are only finitely many types τ such that $\text{depth}(\tau) \leq 1$ and $\text{Par}(\tau) \subseteq A$. But then *AllParSubst*(A) must be finite, too.
2. If Ψ is a parameter substitution such that $\text{depth}(\Psi) = n \geq 1$ and $\text{Par}(\Psi) = \emptyset$, then there exist parameter substitutions Φ_1 and Φ_2 such that
 - (a) $\Phi_1 \in \text{AllParSubst}(\text{dom}(\Psi))$,
 - (b) $\text{depth}(\Phi_2) = n - 1$, and
 - (c) $\Psi = \Phi_1 \circ \Phi_2$.

To prove this, assume $\Psi = [\alpha_1 \mapsto \tau_1, \dots, \alpha_n \mapsto \tau_n]$. For those τ_i such that $\text{depth}(\tau_i) > 1$, we must have $\tau_i = L_i(\sigma_i)$ for some type constructor L_i and some type σ_i with $\text{depth}(\sigma_i) < \text{depth}(\tau_i)$. W.l.o.g. assume that $\text{depth}(\tau_i) \leq 1$ for all $i = 1, \dots, m-1$ and $\text{depth}(\tau_i) > 1$ for all $i = m, \dots, n$. Then define

$$\Phi_1 := [\alpha_1 \mapsto \tau_1, \dots, \alpha_{m-1} \mapsto \tau_{m-1}, \alpha_m \mapsto L_m(\alpha_m), \dots, \alpha_n \mapsto L_n(\alpha_n)] \quad \text{and} \\ \Phi_2 := [\alpha_m \mapsto \sigma_m, \dots, \alpha_n \mapsto \sigma_n].$$

Then the claim is obvious.

3. $\mathcal{I} \approx \{\mathcal{I}\Phi \mid \Phi \in \text{AllParSubst}(\text{Par}(\mathcal{I}))\}$.
Assume $\Phi \models \mathcal{I}$ where w.l.o.g. $\text{dom}(\Phi) \subseteq \text{Par}(\mathcal{I})$. Then the previous property shows that Φ can be written as $\Phi_1 \circ \Phi_2$ where $\Phi_1 \in \text{AllParSubst}(\text{Par}(\mathcal{I}))$. But then $\Phi_2 \models \mathcal{I}\Phi_1$.
Conversely, if $\Psi \models \mathcal{I}\Phi$ for a substitution $\Phi \in \text{AllParSubst}(\text{Par}(\mathcal{I}))$, then $\Phi \circ \Psi \models \mathcal{I}$.
4. If $\Phi \in \text{AllParSubst}(\text{Par}(\mathcal{I}))$, then $\text{depth}(nf(\mathcal{I}\Phi)) \leq \text{depth}(\mathcal{I})$.

Assume $\sigma \preceq \tau$ is an inequation in \mathcal{I} of maximal depth. First, assume $\sigma = K(\sigma')$ and $\tau = L(\tau')$. When going from \mathcal{I} to $nf(\mathcal{I}\Phi)$ this inequation either disappears or it has the form $nf(\sigma'\Phi \preceq \tau'\Phi)$. But the depth of this inequation is not greater than the depth of the original inequation.

Next, $\sigma = \alpha$ for a parameter α and $\tau = L(\tau')$. But then $\sigma\Phi$ must have either of the forms K or $K(\sigma')$. When going from \mathcal{I} to $nf(\mathcal{I}\Phi)$ the inequation $\sigma \preceq \tau$ either disappears or it has the form $nf(\sigma'\Phi \preceq \tau'\Phi)$. Again the depth of

this inequation is not greater than the depth of the original inequation. The remaining cases are similar.

4.1.4 Definition of *Inst*

The function *Inst* transforms a single system of type inequations into an equivalent set of systems of type inequations. It is defined as

$$Inst(\mathcal{I}) := \{nf(\mathcal{I}\Phi) \mid \Phi \in AllParSubst(Par(\mathcal{I})) \wedge nf(\mathcal{I}\Phi) \neq \{\mathbf{false}\}\}.$$

The function *Inst* has the following properties:

1. $Inst(\mathcal{I})$ is finite.
2. $\mathcal{I} \approx Inst(\mathcal{I})$.
3. If $\diamond_k \mathcal{I}$ and $k \geq 1$, then there is a $\mathcal{J} \in Inst(\mathcal{I})$ such that $\diamond_{k-1} \mathcal{J}$.
4. If $\diamond_k \mathcal{J}$ and $\mathcal{J} \in Inst(\mathcal{I})$, then $\diamond_{k+1} \mathcal{I}$.
5. If $\mathcal{J} \in Inst(\mathcal{I})$, then $Par(\mathcal{J}) \subseteq Par(\mathcal{I})$.
6. If $\mathcal{J} \in Inst(\mathcal{I})$, then $depth(\mathcal{J}) \leq depth(\mathcal{I})$.

These properties are immediate consequences of the definition of *Inst* and the properties of the function *AllParSubst*.

4.2 Deciding Type Inequations

We present an algorithm for solving (or refuting) systems of type inequations. The algorithm maintains two sets of systems of inequations. Call these sets \mathfrak{M} and \mathfrak{A} . \mathfrak{M} serves as a memory of systems of type inequations that have already been encountered, while \mathfrak{A} contains systems of type inequations that can be derived from \mathcal{I} by application of the function *Inst*. The algorithm initializes both \mathfrak{M} and \mathfrak{A} to the singleton $\{\mathcal{I}\}$, where \mathcal{I} is the system of type inequations that is to be solved. After this initialization, the algorithm enters a loop. In this loop, we compute $Inst(\mathcal{J})$ for all $\mathcal{J} \in \mathfrak{A}$. Then, we update \mathfrak{A} as follows:

$$\mathfrak{A} := \bigcup \{Inst(\mathcal{J}) \mid \mathcal{J} \in \mathfrak{A}\} - \mathfrak{M}$$

that is, we apply *Inst* to all systems in \mathfrak{A} and we discard those systems that appear already in the memory \mathfrak{M} . If $\emptyset \in \mathfrak{A}$, then \mathcal{I} is solvable and the algorithm halts with success. If \mathfrak{A} becomes empty, the algorithm halts with failure. Otherwise, we update \mathfrak{M} as

$$\mathfrak{M} := \mathfrak{M} \cup \mathfrak{A}$$

and reenter the loop. Figure 1 specifies the algorithm formally.

Lemma 4 (Termination) The algorithm given in Figure 1 terminates.

Proof: For every system of inequations $\mathcal{J} \in \mathfrak{M}$ the number of inequations in \mathcal{J} is less or equal than the number of inequations in \mathcal{I} , $Par(\mathcal{J}) \subseteq Par(\mathcal{I})$, and $depth(\mathcal{J}) \leq depth(\mathcal{I})$. Since the type alphabet is finite, the size of \mathfrak{M} must therefore be bounded.

Now assume the algorithm given in Figure 1 does not terminate. Then the set \mathfrak{A}_{n+1} can never be empty. Therefore, every time the loop is executed, the statement $\mathfrak{M} := \mathfrak{M} \cup \mathfrak{A}_{n+1}$ increases the number of elements of the set \mathfrak{M} . But then the size of \mathfrak{M} would increase beyond every bound. \square

```

Input:  $\mathcal{I}$  % system of type inequations to be solved
 $\mathfrak{M} := \{\mathcal{I}\};$ 
 $\mathfrak{A}_0 := \{\mathcal{I}\};$ 
 $n := 0;$ 
Loop:
   $\mathfrak{A}_{n+1} := \bigcup \{Inst(\mathcal{J}) \mid \mathcal{J} \in \mathfrak{A}_n\} - \mathfrak{M}$ 
  if  $\emptyset \in \mathfrak{A}_{n+1}$  then
    return true;
  end-if;
  if  $\mathfrak{A}_{n+1} = \emptyset$  then
    return false;
  end-if;
   $\mathfrak{M} := \mathfrak{M} \cup \mathfrak{A}_{n+1};$ 
   $n := n + 1;$ 
  goto Loop;

```

Figure 1: An algorithm for deciding solvability of type inequations.

Lemma 5 (Soundness) Assume $n, k \in \mathbb{N}$, $\mathcal{J} \in \mathfrak{A}_n$ and $\diamond_k \mathcal{J}$. Then $\diamond_{k+n} \mathcal{I}$.

Proof: The proof is given by induction on n .

1. $n = 0$: Since $\mathfrak{A}_0 = \{\mathcal{I}\}$ we must have $\mathcal{J} = \mathcal{I}$ and the claim is trivial.
2. $n \rightarrow n + 1$: Assume $\mathcal{J} \in \mathfrak{A}_{n+1}$ with $\diamond_k \mathcal{J}$. Then there is a $\mathcal{K} \in \mathfrak{A}_n$ such that $\mathcal{J} \in Inst(\mathcal{K})$. This implies $\diamond_{k+1} \mathcal{K}$. By i.h. we have $\diamond_{(k+1)+n} \mathcal{I}$. \square

Lemma 6 Assume that $\diamond_k \mathcal{I}$ and k is minimal with this property. Then for all $n \leq k$ there is a $\mathcal{J} \in \mathfrak{A}_n$ such that $\diamond_{k-n} \mathcal{J}$.

Proof: The proof is done by induction on n .

1. $n = 0$: Obvious.
2. $n \rightarrow n + 1$: Assume $\diamond_k \mathcal{I}$ and that k is minimal with this property. By i.h. there is a $\mathcal{J} \in \mathfrak{A}_n$ such that $\diamond_{k-n} \mathcal{J}$. Then there is a $\mathcal{K} \in Inst(\mathcal{J})$ such that $\diamond_{k-n-1} \mathcal{K}$. Assume $\mathcal{K} \in \mathfrak{M}$. Since

$$\mathfrak{M} = \bigcup_{i=1}^n \mathfrak{A}_i,$$

there is an $i \leq n$ such that $\mathcal{K} \in \mathfrak{A}_i$. Therefore Lemma 5 shows $\diamond_{k-n-1+i} \mathcal{I}$. Since $k - n - 1 + i < k$ this contradicts the minimality of k . This shows that the assumption $\mathcal{K} \in \mathfrak{M}$ is wrong and we have $\mathcal{K} \in \mathfrak{A}_{n+1}$. Because of $\diamond_{k-(n+1)} \mathcal{K}$ the proof is complete. \square

Theorem 3 The algorithm given in Figure 1 is correct.

Proof: Assume that \mathcal{I} is solvable. Then $\diamond_n \mathcal{I}$ for some $n \in \mathbb{N}$. By Lemma 6 we find a $\mathcal{J} \in \mathfrak{A}_n$ such that $\diamond_0 \mathcal{J}$ holds. But then $\mathcal{J} = \emptyset$ and the algorithm returns **true**.

Assume now that \mathcal{I} is not solvable. If the algorithm would return **true**, then $\emptyset \in \mathfrak{A}_n$ for some $n \in \mathbb{N}$. Since $\diamond_0 \emptyset$ Lemma 5 would then give $\diamond_n \mathcal{I}$. Therefore the algorithm cannot return **true**. Since it terminates, it must return **false**. \square

5 Conclusion

In this paper we have presented a type system that supports both inclusion polymorphism and parametric polymorphism. We were able to prove that for this type system typability is decidable, provided we use at most unary type constructors. In practice, many interesting type constructors are either nullary or unary. Unary type constructors occur naturally when dealing with container types, e. g. types that are interpreted as sets, lists, or bags. It is convenient to be able to cast, for example, lists to sets. This cannot be done with the type system proposed by Mitchell [Mit84], but is possible with the type system introduced in this paper.

Previously, it has been known that type inference is decidable for a system that restricts inclusion polymorphism to nullary type constructors [FM90, Mit84, Mit91]. On the negative side, Tiuryn and Urzyczyn [TU96] have shown that the type inference problem for second-order types is undecidable. We have shown in this paper, that typability is decidable for type systems with at most unary type constructors. It is still an open question whether typability is decidable in the case of binary type constructors.

Acknowledgement: The authors would like to thank Pawel Urzyczyn for pointing out some technical weaknesses in an earlier version of this paper.

References

- [AM94] Krzysztof R. Apt and Elena Marchiori. Reasoning about Prolog programs: From modes through types to assertions. *Formal Aspects of Computing*, 6A:743–764, 1994.
- [Bei95a] Christoph Beierle. Concepts, implementation, and applications of a typed logic programming language. In Christoph Beierle and Lutz Plümer, editors, *Logic Programming: Formal Methods and Practical Applications*, chapter 5, pages 139–167. Elsevier Science B.V./North-Holland, 1995.
- [Bei95b] Christoph Beierle. Type inferencing for polymorphic order-sorted logic programs. In Leon Sterling, editor, *Proceedings of the 1995 International Conference on Logic Programming*. MIT Press, 1995.
- [FM90] You-Chin Fuh and Prateek Mishra. Type inference with subtypes. *Theoretical Computer Science*, 73(2):155–175, 1990.
- [HL94] Patricia M. Hill and John W. Lloyd. *The Gödel Programming Language*. MIT Press, 1994.
- [HT92] P. M. Hill and R. W. Topor. A semantics for typed logic programs. In Pfenning [Pfe92], pages 1–62.
- [MBL97] Andrew C. Myers, Joseph A. Bank, and Barbara Liskov. Parameterized Types for Java. In *Proceedings of the 24th Symposium on Principles of Programming Languages*, pages 132–145. ACM Press, 1997.
- [Mit84] John C. Mitchell. Coercion and type inference. In *11th Annual ACM Symposium on Principles of Programming Languages*, pages 175–185, 1984.
- [Mit91] John C. Mitchell. Type inference with simple subtypes. *Journal of Functional Programming*, 1:245–285, 1991.
- [OW97] Martin Odersky and Philip Wadler. Pizza into Java: Translating theory into practice. In *Proceedings of the 24th Symposium on Principles of Programming Languages*, pages 146–159. ACM Press, 1997.

- [Pfe92] Frank Pfenning, editor. *Types in Logic Programming*. The MIT Press, 1992.
- [SHC95] Zoltan Somogyi, Fergus J. Henderson, and Thomas Conway. Mercury: an efficient purely declarative logic programming language. In *Proceedings of the Australian Computer Science Conference*, pages 499–512, Glenelg, Australia, February 1995.
- [TU96] Jerzy Tiuryn and Pawel Urzyczyn. The subtyping problem for second-order types is undecidable. In *Proceedings of the IEEE Symposium on Logic in Computer Science (LICS 96)*, pages 74–85, 1996.
- [YFS92] Eyal Yardeni, Thom Frühwirth, and Ehud Shapiro. Polymorphically typed logic programs. In Pfenning [Pfe92], pages 63–90.